14
Measurement of Learning Processes in Pilot Simulation

DEE H. ANDREWS, ROBERT T. NULLMEYER, JUSTINE GOOD, AND PATRICIA C. FITZGERALD

Air Force Research Laboratory

Introduction

There are few human activities as complex as safely piloting an aircraft. All human capabilities and resources are put to the test on a frequent basis in flight, and there are multiple examples of accidents and incidents that have been caused when those capabilities and resources were not sufficient to the task. Examples of capabilities and resources include cognitive, visual, psychoperceptual, kinesthetic, proprioceptive, psychomotor, and social skills. Flight situations call for the pilot to use many of these capabilities and resources simultaneously. Measurement of the performance of pilots requires approaches that take into account all of the complex interactions that occur when so many different human capabilities are used at the same time. As this chapter shows, decades of research and experience have yielded significant advances in the area of pilot performance measurement. Yet, there is still much to be learned about how to measure and analyze this dynamic realm of human behavior.

Researchers and practitioners in the aviation field have a variety of reasons why they need to measure pilot performance accurately and reliably. Key functions for which quality measurements are required include pilot selection, cockpit design, testing, cost estimation, and many others. Much of what we know about pilot measurement has come from functions other

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2008	**Book Chapter**	**00-01-2006 to 00-12-2007**

4. TITLE AND SUBTITLE **Measurement of Learning Processes in Pilot Simulation**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER **62202F**
6. AUTHOR(S) **Dee Andrews; Robert Nullmeyer; Justine Good; Patricia Fitzgerald**	5d. PROJECT NUMBER **1123**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory/RHA,Warfighter Readiness Research Division,6030 South Kent Street,Mesa,AZ,85212-6061**	8. PERFORMING ORGANIZATION REPORT NUMBER **AFRL; AFRL/RHA**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory/RHA, Warfighter Readiness Research Division, 6030 South Kent Street, Mesa, AZ, 85212-6061**	10. SPONSOR/MONITOR'S ACRONYM(S) **AFRL; AFRL/RHA**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) **AFRL-RH-AZ-BC-2008-0001**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Published as Chapter 14 in Baker, E., Dickieson, J., Wulfeck, W., & O'Neil, H.F. (Eds.), Assessment of Problem Solving Using Simulations, pp. 272-288, NY: Lawrence Erlbaum Assoc.

14. ABSTRACT
There are few human activities as complex as safely piloting an aircraft. All human capabilities and resources are put to the test on a frequent basis in flight, and there are multiple examples of accidents and incidents that have been caused when those capabilities and resources were not sufficient to the task. This chapter shows decades of research and experience have yielded significant advances in the area of pilot performance measurement, yet there is still much to be learned about how to measure and analyze this dynamic realm of human behavior. This chapter concentrates on measurement for the training function mainly in simulators.

15. SUBJECT TERMS
Problem solving; Simulations; Performance measurement; Learning processes; Pilot simulation; Pilot performance;

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**16**	

than training, yet training has greatly benefited. In turn, pilot training performance measurement work has produced concepts, methods, and tools that have been beneficial to the other functions. This chapter concentrates on measurement for the training function mainly in simulators.

A Brief Historical Overview of Flight Simulation

The use of flight simulation in pilot training is nearly as old as manned flight itself. The first flight simulator was developed around 1910. This low-technology device, a barrel with short wings that was physically manipulated by the instructor, offered students the opportunity to practice basic flight control (Moroney & Moroney, 1999). During the following two decades, other attempts were made to design flight training devices; the usefulness of these simulators, however, was limited. In 1929, the modern flight simulator was born with the development of the Link trainer, which allowed movement around all three axes of roll, pitch, and yaw (*Link Simulation*, n.d.). Following World War I, the primary focus of performance measurement was on the medical and physiological effects of flight (Moroney & Moroney, 1999). In 1934, the Army Air Corps and the Navy began using flight trainers to assess student pilot performance.

During World War II, the use of the devices for pilot training and selection increased dramatically (*Link Simulation*, n.d). From the end of World War II through the late 1950s, the increasing number and types of aircraft in service and under development prompted the design and deployment of simulators that replicated the characteristics of specific aircraft. During this period, the airlines began using the devices for pilot training. Furthermore, measurement research in simulators expanded to include the development of methods to improve flying safety. In the late 1960s and early 1970s, the advanced processing power of computers led to the development of more sophisticated, high-fidelity devices and the ability to obtain objective performance measures from the simulator (Dickman, 1982). Today, flight simulation is widely used in every aspect of pilot training, from private pilot certification through advanced military distributed mission operations. A number of issues that may affect measurement and performance, however, must be considered.

Issues in the Measurement of Pilot Performance

Several factors may affect the accuracy of performance measurement in simulators, as well as in the operational environment. The evaluation process is complex and depends a great deal on the instructor's judgments of the student's competency. The extent to which the assessment is

objective or subjective is a critical issue. Furthermore, performance in the simulator versus the actual flight environment is qualitatively different, a fact that must also be considered in measurement. Finally, the effects the instructor has on the person evaluated may influence performance, either negatively or positively. These factors are discussed in greater detail.

Subjective versus Objective Measures

The measurement of human performance in aircraft began shortly after the advent of manned flight (Meister, 1999). Performance evaluation occurs on two basic levels. Subjective measures are generally provided by instructors or subject matter experts and assess the performance of the trainee on multiple elements. These data may be qualitative (e.g., comments) or quantitative (e.g., items on a Likert-type scale); however, the metrics often depend on human judgment. Objective measures, on the other hand, consist of specific and well-defined data collected during the training exercise, typically by means of digital computers. The parameters that indicate acceptable performance are empirically based. Although objective data are most desirable, subjective information has been shown to offer rich data that provide a great deal of insight regarding many of the factors that influence human performance (e.g., Nullmeyer, Spiker, Wilson, & Deen, 2003; Spiker & Willis, 2003).

Objective Measures

The development of effective training programs is an iterative process that requires the best performance data available (Vreuls et al., 1975). Although the options for obtaining objective measures of pilot performance in field settings are currently often cost prohibitive, technologies are available to capture this critical information. In actual flight, the military collect objective data through the use of instrument pods (i.e., measurement equipment attached to the wings) that record flight characteristics, weapons information, and flight maneuvers over the course of the mission (Panarisi, 2000).

Because of the costs associated with acquiring the technology required to collect objective performance data in the operational environment, it is more affordable to obtain such data in flight simulators. The current generation of flight simulators affords the collection of a tremendous amount of flight and performance data. Accordingly, training objectives and systematic design principles should be used to ensure that the most appropriate variables are targeted for the performance measurement process (Salas, Milham, & Bowers, 2003).

Subjective Measures

Although objective measures are required for the accurate assessment of performance, the overall picture is incomplete when objective performance data are

relied on exclusively for pilot evaluations. Many of the traditional measures used in the evaluation of pilot performance in operational and simulated environments are subjective, typically including proficiency ratings and instructor comments. Furthermore, in the simulated flight environment, subjective assessments of the students' behavior or performance help to identify the conditions that the instructor selects for subsequent training (Dickman, 1982; Fowlkes, Lane, Salas, Franz, & Oser, 1994; Williams & Thomas, 1984). In efforts to make subjective evaluation more objective, methods have been developed to assess behavior systematically. Event-based methods identify expected behaviors and acceptable responses a priori, and rating scales are designed with specific criteria for all possible levels (Fowlkes et al., 1994; O'Connor, Hormann, Flin, Lodge, & Goeters, 2002). These measures, in conjunction with typical subjective and objective measures, contribute to the accuracy of the overall assessment.

Pilot Behavioral Differences

Of additional interest are pilot behavior differences in the simulator as opposed to actual flight conditions and potential effects on performance. Specifically, the social psychological literature documents instances when performance is altered in the presence of others (e.g., Diaper, 1990; Shivers, 1998; Staal, 2004). Diaper (1990) stated that research participants who were unaware that they were involved in a study performed better than their aware counterparts. In his review of the literature on performance and stress, Staal (2004) reported that performance is enhanced when performing familiar tasks in the presence of others. When engaging in complex tasks that have not yet been well learned, however, performance tends to suffer. Likewise, in his study of employee evaluation and performance, Shivers (1998) concluded that assessment appears to increase apprehension, resulting in negative effects on performance. In operational and simulated environments, the presence of an experimenter or instructor pilot may similarly influence the behavior or performance of the pilot assessed. Accordingly, these factors need to be considered in terms of performance measurement in each environment.

Furthermore, the level of stress experienced by the pilot and the ways in which stressful situations are handled differ in each environment (Staal, 2004). For instance, Wilson, Skelly, and Purvis (1999) reported student pilot heart rates 50% higher in actual flight emergencies in comparison to simulated flight emergencies, suggesting a higher stress level in the actual flight environment. Although studies of this nature suggest the need to be cautious about generalizing simulator study research findings to actual flight, the clear benefits of training in flight simulators are evident in the discussion that follows.

Reliability and Validity Issues in Pilot Training Performance Measurement

To understand reliability and validity challenges in pilot training, it is necessary to understand that such training takes place in different phases and with a range of training media. Initial training, also called ab initio, is conducted with the goal of obtaining a pilot's wings. They are then certified to fly solo. In the military, they then move to training in the actual aircraft they will fly in operational duties (e.g., F-16 fighter, B-1 bomber, C-130 transport). During ab initio and advanced training, they will train using a variety of media. These may consist of paper diagrams, computer-based training, part task training devices with low-to-moderate physical and functional fidelity, sophisticated flight simulators, and finally actual training aircraft. *Physical fidelity* refers to whether the cockpit controls, displays, and out-the-widow scene "look" like the actual aircraft's, and *functional fidelity* refers to whether those items provide the pilot trainee with a "feel" like that of the actual aircraft.

Depending on the phase of pilot training, both reliability and validity of measurement range from fairly straightforward to very complex. For example, if in the initial stages of training we are concerned about a trainee's ability to locate and operate the correct controls in the cockpit, it is not difficult to measure the trainee's performance consistently and validly. Assuming high physical and functional fidelity in the training medium, proper operation of the controls always results in the same functional outcome whether in a training device, a simulator, or an aircraft. In the case of learning the controls and displays and their effect on the aircraft, both reliability and validity can be very high.

However, as pilots move on to more advanced training objectives, such as operating the aircraft in extreme situations (e.g., poor flying conditions or combat flying), it becomes much more difficult to measure and guarantee high reliability and validity. Part of the reason for this difficulty is that, despite some significant advances in the training community's capability to accurately simulate flying events, cues, and sensations (Hawkins, 2002), there are still a variety of areas in which we fall short of attaining complete fidelity with the actual flying environment. For example, although there are promising advances under way in simulating visual out-of-the-cockpit scenes for the pilot, the current state of the art only provides visual fidelity that is less than half the resolution that a pilot can see when flying in the real world. Such decrements in fidelity have a profound effect on our ability to replicate real-world cues validly in a simulated aircraft and to measure reliably a pilot's reaction in a complex setting. In the case of visual cues, it is possible that the pilot might do something differently in the actual aircraft because there is not a full range of visual cues.

The reliability and validity issue is particularly pronounced in training combat skills. For example, take the case of fighter pilots training to fight against other fighter pilots in air-to-air engagements. The various standard fighter combat maneuvers (e.g., barrel roles, Immelman turn, low yo-yo, drag) can be taught and practiced with fairly well-defined measures of performance in a training setting. It is difficult to judge how well the maneuvers were performed in actual combat because the recording of the maneuvers either in the aircraft or from a ground monitoring station can be difficult in combat situations and because pilots may have to perform the maneuver much differently from the accepted standard maneuver because of the conditions of the engagement.

Combat also presents the most difficult cognitive performance tasks because of the great novelty and variety a pilot faces in various combat settings. It is difficult to establish standards for reliability in such situations because of the need for flexibility in the way a combat pilot assesses the situation and develops a solution. In turn, the validity of various measures of combat performance is difficult to establish because the training goal is to provide the trainee with a reasonable sampling of combat situations. Because of the subjective nature of the "goodness" of a particular solution to an actual combat setting, the validity may or may not be high when transfer from the training setting to the real-world setting is considered. Indeed, much of pilot performance measurement is derived from expert judgments of that performance so that even in relatively straightforward areas like cockpit procedures training each expert might have variations in how consistently and validly those subjective measurements are made compared to their fellow experts and even with themselves over time.

The remainder of the chapter explores pilot performance measurement issues in two related but different flight domains: ab initio training (i.e., beginning undergraduate pilot training) and wide-body aircraft flight training.

Performance Measurement In Simulation-Based Undergraduate Pilot Training

The U.S. Air Force (USAF) trains officers to become pilots through a 52-week program currently known as Joint (meaning shared with the U.S. Navy and U.S. Army) Specialized Undergraduate Pilot Training. The course objectives throughout the phases of Joint Specialized Undergraduate Pilot Training remain the same: to qualify graduates for the aeronautical rating of pilot and for follow-on phases of training and for future responsibilities as military officers and leaders. This training includes flying training to teach the principles and techniques used in operating high-speed jet aircraft; ground training to supplement and reinforce flying training; and officer development training to

strengthen the graduate's leadership skills, officer qualities, and understanding of the role of the military pilot as an officer and supervisor.

The basic methodology for USAF training is first to instruct the concepts in an academic training environment, demonstrate the concept to the student in a training medium either on the ground or in the air, then provide the student an opportunity to practice the concept. Phase I training begins with academic training. Once the students have a sufficient amount of "book knowledge," they are introduced to some basic flying concepts in the simulator. Measurement of their academic learning is typically straightforward and possesses both high reliability and high validity. When the students complete Phase I, they continue to the next phase of training in a specific aircraft.

Newer simulator systems for undergraduate pilot training, such as for the T-6 (a single-engine, tandem-cockpit propeller aircraft), offer ways to record and measure student performance electronically during a simulator sortie (a *sortie* is one flight regardless of whether it is in the simulator or the actual aircraft). However, the majority of the grading and evaluation is done by instructor observation. In the older training devices, such as the T-1 (a jet engine, business jet-like trainer) and T-38A (a twin-engine, high-performance jet trainer) simulators, it is possible to record and even sometimes document the aircraft's flight path and performance electronically during certain maneuvers (e.g., an instrument approach). Although the system is not readily used by instructors for evaluation because they do not find the performance measurement system user friendly, it may be used to debrief the particular maneuver or the overall sortie if necessary. One of the more commonly used capacities of the simulator is the ability to have the student perform a maneuver, and if the student incorrectly performs the maneuver, the simulator can be stopped, instruction can be offered to the student, and the student can be given another opportunity to perform the maneuver. In many cases, some simulators have the ability to record the performance parameters and replay the exact student performance so the student can watch what he or she was doing to the aircraft during the maneuver. This helps the student recognize errors and learn from them. It also provides an excellent opportunity for the instructor to show the student exactly what was going on at the time and have the students' full attention during the instruction.

For undergraduate pilot training, the grading standards for both flying and simulator training are defined as follows:

No grade: The maneuver is demonstrated by the instructor pilot but not practiced by the student.

Unable to accomplish: The student is unsafe or lacks sufficient knowledge, skill, or ability to perform the operation, maneuver, or task.

Fair: The student performs the operation, maneuver, or task safely but has limited proficiency. Deviations occur that detract from performance.
Good: The student performs the operation, maneuver, or task satisfactorily. Deviations occur that are recognized and corrected in a timely manner.
Excellent: The student performs the operation, maneuver, or task correctly, efficiently, and skillfully.

Although there are relatively clear descriptions of the standards used to judge which of the rating categories should be assigned, considerable latitude is given to the instructor in making the subjective evaluations.

By the nature of flying, during every sortie the students are presented with a number of variables and problems they must solve to complete the mission safely and successfully. Sometimes, these issues are known and anticipated; however, many times they are unexpected. Some of the most critical areas of instruction and most difficult tasks/concepts for the students to develop and grasp involve a significant amount of problem solving: risk management/decision making, task management, and situational awareness (SA). For example, in the T-1 undergraduate syllabus, for risk management/decision making (T-1, 2003, p. 16), students are expected to:

a. Identify probable contingencies and alternatives
b. Gather available data before arriving at final decision
c. Encourage crew participation in the decision making process
d. Clearly state decisions to the crew
e. Provide rationale for decisions

Similar to this is task management, for which the student is expected to prioritize multiple tasks correctly and use all available resources to manage workload. Clearly, measuring and evaluating proficiency in all these skills is difficult.

SA is the skill area students typically have the most difficulty developing. These difficulties often lead to a large percentage of unsatisfactory performances on evaluation sorties. Although each aircraft may necessitate a somewhat different definition of SA, essentially the SA concept can be defined as it is in the T-1 syllabus (Joint Specialized Undergraduate Pilot Training, 2003, p. 16):

a. *Awareness:* Keep track of what is happening on the ground, in the air, and with other crew members. Cope with any subsequent mission impact as a result of these happenings.
b. *Flexibility:* Cope with rapidly changing situations or conditions, inflight or on the ground, and adjust mission as needed to obtain desired objectives.
c. *Capacity:* Cognizant of the awareness level of self and other crew members and acts to maintain a high level of SA for all.

The students are expected to demonstrate the ability to maintain awareness and minimize the effects of adverse factors on the crew. These skills are required in the student pilot, who must maintain and recognize the SA of the other members of the flight. The bottom line of SA is that the pilot never allows the crew to exceed their capability to fly safely. Again, for an instructor to measure and evaluate such a complex cognitive function as SA in a simulator or in the aircraft takes a gifted instructor because of the task's subjectivity.

Crew Performance Measurement in Wide-Body Flight Simulators

Wide-body aircraft are generally considered to be any aircraft that have multiple crew members. The wide-body aircraft is larger than a small training aircraft or fighter aircraft. Examples of wide-body aircraft are commercial jet liners, civilian and military transport aircraft, bombers, and military tankers. Pilot measurement issues in wide-body aircraft can be more complex than in non-wide-body aircraft because of the multiple-person crew.

The backbone of student performance measurement in military wide-body simulator training programs is instructor observation of student behaviors, including associated impacts on the simulated training environment. Automated performance measurement capabilities may augment these observations, serving at least two purposes: (a) to enhance the instructor's awareness of student behavior in the instructional environment and (b) to improve the quality of feedback given to students. There is long-standing interest in using performance measurement capabilities for a third purpose: to support competency-based progression through simulator training experiences. However, students are given fairly fixed sequences of experiences in most simulator training programs today. Limited availability of simulators (with the simulator sometimes a one-of-a-kind device) often results in limited scheduling flexibility. In addition, tailoring instruction to the needs of one individual in a multiperson crew is complicated by impacts on the other crew positions.

Instructor-Based Performance Measurement

Instructor observations remain the primary inputs both for posttraining mission debriefing and for documenting the adequacy of student progress in student records. Observations are recorded in several ways. Using C-130 training as an example, instructors fill out an Aircrew Training Progress

Record after each simulator mission. This form provides a set of required proficiency levels for task-based training events such as airdrop checklist, simulated engine failure, night vision device operations, and so forth. Performance and knowledge are rated using 4-point knowledge and skill scales. The second method for documenting student performance is instructor comments provided for each simulator mission on a separate Training Comments Record. The comments are unstructured and are not necessarily tied to the required items covered in the Progress Record. Instructors may laud exemplary performance or describe deficiencies and can use the Training Comments Record as a teaching or debriefing aid. Instructor comments have proven to be good sources of insight concerning student strengths and weaknesses in both Navy and Air Force applications.

Spiker, Berkman, and Hunt (2002) analyzed S-3B aircraft student training records (two-person crews) from Navy familiarization training. Each instructor comment was assigned to a category and subcategory within a comprehensive Aircrew Proficiency Classification Framework. Example categories (and subcategories) included perception (cue detection, perceptual illusion); knowledge (of systems, operating limits); procedural (checklist, standard operating procedures); and so forth. The remaining categories were aircraft handling, task management, communication, crew coordination, attitude, decision making, situation awareness, thinking patterns, mission assessment, and emergency procedures. Frequencies of positive and negative comments proved an effective way to pinpoint strong areas of S-3B training effectiveness (crew backup) and weak areas that represent opportunities to improve instruction (communication discipline, attitude awareness).

Spiker and Willis (2003) applied the S-3B taxonomy to review C-130 instructor comments. Of some note, decision making and risk assessment were virtually never mentioned in C-130 student records, yet they were two of the leading factors in C-130 mishaps (Nullmeyer et al., 2003). Subsequent discussions with instructors revealed that these skill areas were no longer emphasized in simulator and flight training, which may explain their prominence in mishap reports. For both S-3 and C-130 instruction, instructor comments helped identify friction points in the training process that had gone undetected. We view instructor comments to be powerful but typically untapped data that can be used to gauge both student proficiency and training effectiveness.

The nature of comments recorded by C-130 simulator instructors may have implications for automated performance measurement. Comments from C-130 mission qualification simulator training were divided into two groups: (a) task-related skills such as task execution, procedures, checklist accomplishment, and aircraft handling; and (b) more cognitive skills such

as crew coordination, SA, and mission planning. Overall, 39% of comments were task oriented, and 61% pertained to more cognitive aspects of student performance such as crew coordination, communication, SA, and mission planning/evaluation (data obtained from Spiker & Willis, 2003). These proportions varied across crew positions. For student aircraft commanders, most simulator instructor comments pertained to procedures and tasks or aircraft handling. For loadmasters, comments were evenly distributed across procedures and more cognitive skills. For the remaining crew positions, the largest proportions of comments addressed cognitive skills. These data are summarized in Table 14.1. It is clear that a comprehensive performance measurement system must address cognitive skills.

The skills included in the bottom row of Table 14.1 overlap considerably with traditional crew resource management (CRM) skills. For Air Force aviators, CRM is defined in terms of six skill areas: mission planning, SA, communication, risk assessment/decision making, task management, and crew coordination/flight integrity. Researchers in all military services have successfully used behaviorally anchored rating scales to capture and quantify CRM skills. Using this measurement approach, CRM skill ratings have been strong predictors of mission performance. In several studies, CRM skill levels for experienced crews were measured during annual simulator refresher training using 5-point behaviorally anchored rating scales. These scales addressed specific aspects of each CRM skill category. Both the specific aspects to be measured and the behavioral anchors that exemplify the points were populated with inputs from platform-specific subject matter experts. Thompson, Tourville, Spiker, and Nullmeyer (1999) and Nullmeyer and Spiker (2003) reported equally strong CRM/mission performance correlations in simulator training for MH-53J and MC-130P crews, respectively. In each of these research studies, subject matter experts used paper-based forms to capture CRM skills. This measurement approach has now been modified to support continuing CRM data collection during operational simulator training for MC-130P student crews to

Table 14.1 Frequency (Percentage) of C-130 Simulator Instructor Comments by Crew Position

Skill area	Aircraft Commander	Copilot	Navigator	Flight Engineer	Loadmaster
Task-related skills	164 (58%)	72 (38%)	42 (20%)	85 (32%)	110 (52%)
Cognitive skills	124 (43%)	118 (62%)	170 (80%)	179 (68%)	103 (48%)

guide subsequent training to address areas of greatest need (Thompson et al., 1999).

Automated Performance Measurement

Performance monitoring and reporting capabilities are included as instructional features in most high-fidelity simulators. Many early systems (delivered in the 1980s) captured aircraft system status data and flight parameters like air speed and altitude. Polzella, Hubbard, Brown, and McLean (1987) surveyed over 100 Air Force C-130, H-53, E-3A, and B-52 simulator instructors concerning frequency of use and value of parameter/procedure monitoring tools designed to enhance instructor awareness in simulators. Instructors who were located at an external instructor/operator station (IOS) reported that performance and procedure monitoring capabilities were frequently used and had moderate-to-high training value. Instructors who were colocated with students could observe targeted behaviors "over the shoulder" directly. These instructors expressed a strong preference for direct monitoring of student performance to include cockpit displays, crew interactions, and the out-the-window scene. Utility and utilization ratings for monitoring tools by these instructors were significantly lower but still moderately positive.

Polzella and his colleagues (1987) also addressed the utility and utilization of enhanced student feedback capabilities. Features corresponding to this function included record/playback and hard copy printouts of flight parameters, which were available on the majority of devices. VHS recorders were also common to capture crew interactions. Training value and utilization ratings for these capabilities were both generally low. Many instructors reported that performance retrieval with these features was time consuming, unreliable, and difficult. In addition, products were often difficult to interpret.

MH-53J helicopter simulator instructors were surveyed to identify opportunities for improving instructor/simulator interfaces (Nullmeyer, Cicero, Spiker, Tourville, & Thompson, 1998). These instructors are colocated in the simulator with their students and yet indicated a high level of interest in monitoring capabilities that added to their awareness of specific aspects of the training environment, especially displays that provided knowledge of the electronic combat environment and accurate position information relative to salient objects such as terrain, threats, cultural features, and planned routes and way points. Instructors viewed performance-capturing capabilities to improve feedback as "nice to have," but they consistently gave a higher priority to increased awareness. This may have been influenced by their experience with automated performance

measurement capabilities that they described as not providing the information they would use to enhance debriefings.

Automated simulator performance measurement technology is advancing along at least two fronts. The first involves the technology itself. Much of the limited enthusiasm in early simulator instructor surveys may have been the product of equipment such as VHS recorders that made data retrieval so cumbersome and time consuming that instructors often viewed such features as more disruptive than beneficial. Again using the C-130 community as an example, automated data visualization and analysis capabilities have been added to the C-130 full-mission simulators. Flight data are digitally recorded and can be displayed in several forms, ranging from graphs and navigational charts to synchronized video and detailed three-dimensional graphic animations of the flight. Event markers and other retrieval tools resolve many of the problems reported with earlier technologies. Event markers are digital time stamps that the instructor can make in the simulator's data archiving system so that, after the simulator sortie, the instructor can quickly return the student to a critical part of the sortie for replay and remediation if necessary.

The second area of advancement is broadening the types of data that are captured. Military training researchers are finding that aircraft system status and aircraft position data, although important, do not suffice as stand-alone measures of crew performance; other factors are emerging as essential elements. One major emerging factor is mission preparation. Bergondy, Fowlkes, Gualtieri, and Salas (1988) found that well over half of debriefing items in Navy Air Wing Integration Training addressed mission planning and briefing, and only 42% addressed execution issues. Spiker, Nullmeyer, and Tourville (2001) found that MC-130P crew interactions during planning and briefing for a simulator mission accounted for over 60% of variance (r = .78) in independent expert ratings of mission performance, a relationship that was also reported by Thompson et al. (1999) for rotary wing crews (76%, r = .87). Consistent with these data, the capability to add crew plan information into the performance monitoring capability of MH-53 simulators emerged as a highly desirable feature. Clearly, crew interaction skills like communication and crew coordination will also need to be addressed.

These needs are reflected in common simulator IOS requirements that were established as part of the Navy Aviation Simulation Master Plan (Walwanis Nelson, Smith, Owens, & Bergondy-Wilhelm, 2003). An IOS provides the instructor and simulator's operator with a variety of methods for controlling the simulator sortie and for recording the trainee's performance. In the Navy Aviation Simulator Master Plan, performance measurement is reaffirmed as a major function of the IOS. The Navy vision

incorporates both automatic and manual measurement. Automatic recording capabilities would be based on trigger events and give instructors more references to support the debriefing. Manual measurement would be supported by the capability to insert event markers that allow instructors to highlight particular moments in the scenario and retrieve the desired information quickly and easily. Of 26 possible IOS requirements, only data-recording capabilities and bird's-eye view playback were identified by all platforms, indicating strong instructor support for these functions. In simulator terms, a bird's-eye view allows the instructor and trainee to look at a map of the simulated terrain over which the trainee has flown the training sortie.

Conclusion

As shown in this chapter, the aviation community's ability to measure the performance of pilot trainees accurately and validly has grown tremendously since the early days of flight. The use of advanced simulators for undergraduate and advanced pilots has opened many doors to better performance measurement. The days of requiring an instructor pilot to make all of the judgments about trainee pilots based solely on their own subjective observations are now finished. However, even with all of the automated simulation-based performance measurement tools described in this chapter, it will always be up to an experienced instructor pilot to make the final instructional and evaluative decisions about the trainees.

Acknowledgment

The opinions expressed in this chapter are those of the authors and do not necessarily represent the official views or policies of the Department of Defense or the Department of the Air Force.

References

Bergondy, M., Fowlkes, J., Gualtieri, J. & Salas, E. (1998, November). Key team competencies for Navy air wings: A case study. In *Proceedings of the 19th Interservice/Industry Training Systems and Education Conference*. Arlington, VA: National Training Systems Association.

Diaper, G. (1990). The Hawthorne effect: A fresh examination. *Educational Studies, 16*, 261–267.

Dickman, J. L. (1982, November). Automated performance measurement: An overview. In *Proceedings of the 4th Interservice/Industry Training Equipment Conference* (pp. 153–165). Arlington, VA: National Security and Industrial Association.

Fowlkes, J. E., Lane, N. E., Salas, E., Franz, T., & Oser, R. (1994). Improving the measurement of team performance: The TARGETs methodology. *Military Psychology, 6*, 47–61.

Hawkins, F. H. (2002). *Human factors in flight.* Aldershot, UK: Ashgate.

Link simulation and training setting the standard for over 75 years. (n.d.) Retrieved January 6, 2006, from http://www.link.com/history.html

Meister, D. (1999). Measurement in aviation systems. In D. J. Garland, J. A. Wise, and V. D. Hopkin (Eds.), *Handbook of aviation human factors* (pp. 34–49). Mahwah, NJ: Erlbaum.

Moroney, W. F., & Moroney, B. W. (1999). Flight simulation. In D. J. Garland, J. A. Wise, and V. D. Hopkin (Eds.), *Handbook of aviation human factors* (pp. 355–388). Mahwah, NJ: Erlbaum.

Nullmeyer, R. T., Cicero, G. D., Spiker, V. A., Tourville, S. J., and Thompson, J. S. (1998, November). Improving instructor operator stations to enhance electronic warfare training. In *Proceedings of the 20th Interservice/Industry Training Systems and Education Conference.* Arlington, VA: National Training Systems Association.

Nullmeyer, R. T., & Spiker, V. A. (2003). The importance of crew resource management in MC-130P mission performance: Implications for training evaluation. *Military Psychology, 15*, 77–96.

Nullmeyer, R., Spiker, V. A., Wilson, D., and Deen, G. (2003, November). Key crew resource management behaviors underlying C-130 aircrew performance. In *Proceedings of the Interservice/Industry Training Systems and Education Conference.* Arlington, VA: National Training Systems Association.

O'Connor, P., Hormann, H. J., Flin, R., Lodge, M., &Goeters, K. M. (2002). Developing a method for evaluating crew resource management skills: A European perspective. *International Journal of Aviation Psychology, 12*, 263–285.

Panarisi, M. T. (2000). A comparative analysis of internal and external solutions to provide air combat maneuvering instrumentation (ACMI) functionality. School of Advanced Air Power Studies, Air University, Maxwell Air Force Base, AL. Retrieved January 6, 2006 from Defense Technical Information Center Scientific and Technical Information Network Web site: http://stinet.dtic.mil/

Polzella, D. J., Hubbard, D. C., Brown, J. E., and McLean, H. C. (1987). *Aircrew training devices: Utility and utilization of advanced instructional features (phase IV summary report)* (AFHRL-TR-87-21, AD-A188 418). Williams AFB, AZ: Operations Training Division. Project 1123, Contract F3315-84-C-0066, University of Dayton Research Institute.

Salas, E., Milham, L. M., & Bowers, C. A. (2003). Training evaluation in the military: Misconceptions, opportunities, and challenges. *Military Psychology, 15*, 3–16.

Shivers, C. H. (1998). Halos, horns and Hawthorne: Potential flaws in the evaluation process. *Professional Safety, 43*(3), 38–41.

Spiker, V. A., Berkman, A. A., & Hunt, S. (2002, November). *Identifying human factors trends in VS-41 training records* (ASI Tech. Rep. 1450-1). Santa Barbara, CA: Anacapa Sciences.

Spiker, V. A., Nullmeyer, R. T., and Tourville, S. J. (2001, November). Relationship between mission preparation and performance during combat mission training. *Proceedings of the 23nd Interservice/Industry Training Systems and Education Conference.* Arlington, VA: National Training Systems Association.

Spiker, V. A., & Willis, P. (2003). *Analysis of C-130 training records to identify critical CRM behaviors* (ASI Tech. Rep. 1392-2). Santa Barbara, CA: Anacapa Sciences.

Staal, M. A. (2004). *Stress, cognition, and human performance: A literature review and conceptual framework* (NASA Tech. Rep. NASA/TM-2004-212824). Ames Research Center, Moffat Field, CA. Retrieved January 5, 2006, from http://humanfactors.arc.nasa.gov/web/publications/20051028105746_IH-054%20Staal.pdf

T-1 Joint Specialized Undergraduate Pilot Training/Fixed-Wing Transition Syllabus P-V4A-G. (2003, December). Randolph AFB, TX: Headquarters Air Education and Training Command.

Thompson, J. S., Tourville, S. J., Spiker, V. A., and Nullmeyer, R. T. (1999, November). Crew resource management and mission performance during MH-53J combat mission training. *Proceedings of the 21st Interservice/Industry Training Systems and Education Conference.* Arlington, VA: National Training Systems Association.

Vreuls, D., Wooldridge, A. L., Obermeyer, R. W., Johnson, R. M., Norman, D. A., & Goldstein, I. (1975). *Development and evaluation of trainee performance measures in an automated instrument flight maneuvers trainer* (NAVTRAEQUIPCEN 74-C-0063-1). Naval Training Equipment Center, Orlando, FL: Human Factors Laboratory. Retrieved January 6, 2006 from Defense Technical Information Center Scientific and Technical Information Network Web site: http://stinet.dtic.mil/

Walwanis Nelson, M. M., Smith, D. G., Owens, J. M., & Bergondy-Wilhelm, M. L. (2003, November). A common instructor operator station framework: Enhanced usability and instructional capabilities. *Proceedings of the Interservice/Industry Training Systems and Education Conference.* Arlington, VA: National Training Systems Association.

Williams, A. M., & Thomas, G. S. (1984, November). A research tool to improve the effectiveness of performance measurement within the IOS. In *Proceedings of the Sixth Interservice/Industry Training Equipment Conference* (pp. 117–120). Washington, DC: National Security and Industrial Association.

Wilson, G. F., Skelly, J., & Purvis, B. (1999). Reactions to emergency situations in actual and simulated flight. *Human Performance in Extreme Environments,* 4(2), 34–45.